

# Uniform in Bandwidth consistency of Kernel-Type Estimators of Shannon's Entropy

Salim BOUZEBDA & Issam ELHATTAB

*L.S.T.A., Université de Paris 6. 175, rue du Chevaleret, 8ème étage, bâtiment A,*

*75013 PARIS FRANCE.*

---

**Abstract.** We establish uniform-in-bandwidth consistency for kernel-type estimators of the differential entropy. Our proofs rely on the methods of Einmahl and Mason (2005).

*Keywords:* Differential entropy; Kernel density estimation; Consistency; Uniform in bandwidth.

---

## 1 Introduction - Main Results

Let  $\mathbf{X}_1, \dots, \mathbf{X}_n$  be independent random replicæ of a random vector  $\mathbf{X} \in \mathbb{R}^d$  with distribution function  $\mathbb{F}(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x})$  for  $\mathbf{x} \in \mathbb{R}^d$ , with  $d \geq 1$ . We set here  $\mathbf{X} = (X_1, \dots, X_d) \leq \mathbf{x} = (x_1, \dots, x_d)$  whenever  $X_i \leq x_i$ , for all  $i = 1, \dots, d$ . We assume that the distribution function  $\mathbb{F}(\cdot)$  has a density  $f(\cdot)$  (w.r. to Lebesgue measure in  $\mathbb{R}^d$ ). The differential entropy of  $f(\cdot)$  is then given by

$$H(f) = - \int_{\mathbb{R}^d} f(\mathbf{x}) \log(f(\mathbf{x})) d\mathbf{x}, \quad (1)$$

whenever this integral is meaningful, and where  $d\mathbf{x}$  denotes Lebesgue measure in  $\mathbb{R}^d$ . The notion of differential entropy was essentially introduced by Shannon (1948). We refer to Cover and Thomas (2006) (see their Chapter 8), and the references therein, for details. Because of numerous applications, the problem of estimating  $H(f)$  has been the subject of considerable interest in the last decades (refer to Beirlant *et al.* (1997) and the references therein). The main purpose of the present article is to establish consistency and to provide asymptotic confidence intervals for the entropy functional  $H(f)$ , based on kernel-type functional estimators.

Below, we will work under the following assumptions on  $f(\cdot)$ .

**(F.1)**  $H(f)$  is properly defined by the integral (1), in the sense that

$$|H(f)| < \infty; \quad (2)$$

**(F.2)**  $f(\cdot)$  is bounded and strictly positive on  $\mathbb{R}^d$ .

We refer to Györfi and van der Meulen (1991) for conditions characterizing (2) in terms of  $f(\cdot)$ . To define our entropy estimator we define, in a first step, a kernel density estimator. Towards this aim, we introduce a measurable function  $K(\cdot)$  fulfilling the conditions

**(K1)**  $K(\cdot)$  is of bounded variation on  $\mathbb{R}^d$ ;

**(K2)**  $K(\cdot)$  is right continuous on  $\mathbb{R}^d$ , i.e.,

$$K(t_1, \dots, t_d) = \lim_{\varepsilon_1 \downarrow 0, \dots, \varepsilon_d \downarrow 0} K(t_1 + \varepsilon_1, \dots, t_d + \varepsilon_d);$$

**(K3)**  $\|K\|_\infty := \sup_{\mathbf{x} \in \mathbb{R}^d} |K(\mathbf{x})| =: \kappa < \infty$ ;

**(K4)**  $\int_{\mathbb{R}^d} K(\mathbf{s}) d\mathbf{s} = 1$ .

We then make use of an Akaike-Parzen-Rosenblatt (refer to Akaike (1954), Parzen (1962) and Rosenblatt (1956)) kernel estimator of  $f(\cdot)$ , defined as follows. Given a bandwidth sequence  $0 < h_n \leq 1$ , we estimate  $f(\mathbf{x})$  by

$$\hat{f}_{n,h_n}(\mathbf{x}) = (nh_n)^{-1} \sum_{i=1}^n K((\mathbf{x} - \mathbf{X}_i)/h_n^{1/d}), \quad \text{for } \mathbf{x} \in \mathbb{R}^d. \quad (3)$$

In a second step, given  $\hat{f}_{n,h_n}(\cdot)$ , we estimate  $H(f)$  by setting

$$H_{n,h_n,\beta}(f) = - \int_{A_{n,\beta}} \hat{f}_{n,h_n}(\mathbf{x}) \log(\hat{f}_{n,h_n}(\mathbf{x})) d\mathbf{x}, \quad (4)$$

where  $A_{n,\beta} := \{\mathbf{x} : \hat{f}_{n,h_n}(\mathbf{x}) \geq (\log_+ n)^{-\beta}\}$  and  $\beta \in (0, 1/4)$  is a specified constant. Here, we set  $\log_+(u) = \log(u \vee e)$  for  $u \in \mathbb{R}$ .

The limiting behavior of  $\hat{f}_{n,h_n}(\cdot)$ , for appropriate choices of the bandwidth  $h_n$ , has been extensively investigated in the literature (refer to Bosq and Lecoutre (1987), Devroye and Györfi (1985) and Devroye and Lugosi (2001)). In particular, under our assumptions, the condition that  $h_n \rightarrow 0$  together with  $nh_n \rightarrow \infty$  is necessary and sufficient for the convergence in probability of  $\hat{f}_{n,h_n}(\mathbf{x}) \rightarrow f(\mathbf{x})$ , independently of  $\mathbf{x} \in \mathbb{R}^d$  and  $f(\cdot)$ . Recently, Deheuvels (2000), Einmahl and Mason (2000), Deheuvels and Mason (2004), Einmahl and Mason (2005) and Dony and Einmahl (2006) established uniform consistency results for such estimators, where  $h_n$  varies within suitably chosen intervals indexed by  $n$ . In the present paper we will use their methods to establish convergence results for  $H_{n,h_n,\beta}(f)$ .

Set

$$\hat{\mathbb{E}}H_{n,h_n,\beta}(f) = - \int_{A_{n,\beta}} \mathbb{E}\hat{f}_{n,h_n}(\mathbf{x}) \log(\mathbb{E}\hat{f}_{n,h_n}(\mathbf{x})) d\mathbf{x}.$$

Select a sequence of constants  $\{b_n : n \geq 1\}$  such that  $b_n \downarrow 0$  and  $nb_n/(\log n)^{1+4\beta} \rightarrow \infty$ . Our main result is as follows.

**Theorem 1.1** *Let  $K(\cdot)$  satisfy (K1)-(K4), and let  $f(\cdot)$  fulfill (F1)-(F2). Then for each  $\beta \in (0, 1/4)$ , there exists a function  $\Upsilon(c)$  of  $c > 0$ , such that, for each  $c > 0$  satisfying  $0 < cn^{-1}(\log n)^{1+4\beta} < b_n \leq 1$  with  $b_n \downarrow 0$ , we have*

$$\limsup_{n \rightarrow \infty} \sup_{cn^{-1}(\log n)^{1+4\beta} \leq h \leq b_n} \frac{\sqrt{nh}|H_{n,h,\beta}(f) - \hat{\mathbb{E}}H_{n,h,\beta}(f)|}{\sqrt{\{\log n\}^{4\beta}(\log(1/h) \vee \log \log n)}} \leq \Upsilon(c) \text{ a.s.}$$

Let  $\{a_n : n \geq 1\}$  be a sequence of constants such that  $0 < a_n < b_n < 1$ , together with  $na_n/(\log n)^{1+4\beta} \rightarrow \infty$ . An application of Theorem 1.1 shows that, with probability 1,

$$\sup_{a_n \leq h \leq b_n} |H_{n,h,\beta}(f) - \hat{\mathbb{E}}H_{n,h,\beta}(f)| = O \left( \sqrt{\frac{\{\log n\}^{4\beta}(\log(1/a_n) \vee \log \log n)}{na_n}} \right).$$

This, in turn, implies that

$$\lim_{n \rightarrow \infty} \sup_{a_n \leq h \leq b_n} |H_{n,h,\beta}(f) - \hat{\mathbb{E}}H_{n,h,\beta}(f)| = 0 \text{ a.s.} \quad (5)$$

Thus we have the following corollary of Theorem 1.1

**Corollary 1.2** *Let  $K(\cdot)$  satisfy (K1)-(K4), and let  $f(\cdot)$  be a uniformly Lipschitz continuous, and strictly positive density, on  $\mathbb{R}^d$ , fulfilling (F1). Then for any  $\beta \in (0, 1/4)$ , and for each pair of sequences  $0 < a_n < b_n \leq 1$  with  $b_n \downarrow 0$  and  $na_n/(\log n)^{1+4\beta} \rightarrow \infty$ , we have*

$$\sup_{a_n \leq h \leq b_n} |H_{n,h,\beta}(f) - H(f)| \rightarrow 0. \quad (6)$$

We note that the main problem in using entropy estimates such as (4) is to choose properly  $h_n$ . The uniform in bandwidth consistency result given in (6) shows that any choice of  $h$  between  $a_n$  and  $b_n$  ensures the consistency of  $H_{n,h,\beta}(f)$ .

## 2 Proofs

**Proof of Theorem 1.1.** We first decompose  $H_{n,h_n,\beta}(f) - \hat{\mathbb{E}}H_{n,h_n,\beta}(f)$  into the sum of two components, by writing

$$\begin{aligned} H_{n,h_n,\beta}(f) - \hat{\mathbb{E}}H_{n,h_n,\beta}(f) &= - \int_{A_{n,\beta}} \hat{f}_{n,h_n}(\mathbf{x}) \log(\hat{f}_{n,h_n}(\mathbf{x})) d\mathbf{x} + \\ &\quad + \int_{A_{n,\beta}} \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) \log(\mathbb{E} \hat{f}_{n,h_n}(\mathbf{x})) d\mathbf{x} \\ &= - \int_{A_{n,\beta}} \left\{ \log \hat{f}_{n,h_n}(\mathbf{x}) - \log \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) \right\} \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) d\mathbf{x} \\ &\quad - \int_{A_{n,\beta}} \left\{ \hat{f}_{n,h_n}(\mathbf{x}) - \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) \right\} \log \hat{f}_{n,h_n}(\mathbf{x}) d\mathbf{x} \\ &:= \Delta_{1,n,h_n,\beta} + \Delta_{2,n,h_n,\beta}. \end{aligned} \quad (7)$$

We observe that for all  $z > 0$ ,  $|\log z| \leq \left| \frac{1}{z} - 1 \right| + |z - 1|$ . Therefore, we get

$$\begin{aligned}
|\log \hat{f}_{n,h_n}(\mathbf{x}) - \log \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x})| &= \left| \log \frac{\hat{f}_{n,h_n}(\mathbf{x})}{\mathbb{E} \hat{f}_{n,h_n}(\mathbf{x})} \right| \\
&\leq \left| \frac{\mathbb{E} \hat{f}_{n,h_n}(\mathbf{x})}{\hat{f}_{n,h_n}(\mathbf{x})} - 1 \right| + \left| \frac{\hat{f}_{n,h_n}(\mathbf{x})}{\mathbb{E} \hat{f}_{n,h_n}(\mathbf{x})} - 1 \right| \\
&= \frac{|\mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) - \hat{f}_{n,h_n}(\mathbf{x})|}{\hat{f}_{n,h_n}(\mathbf{x})} + \frac{|\hat{f}_{n,h_n}(\mathbf{x}) - \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x})|}{\mathbb{E} \hat{f}_{n,h_n}(\mathbf{x})}.
\end{aligned}$$

Recalling that  $A_{n,\beta} := \{\mathbf{x} : \hat{f}_{n,h_n}(\mathbf{x}) \geq (\log n)^{-\beta}\}$ , we readily obtain from these relations that, for any  $\mathbf{x} \in A_{n,\beta}$ ,

$$|\log \hat{f}_{n,h_n}(\mathbf{x}) - \log \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x})| \leq \frac{2}{(\log n)^{-\beta}} |\hat{f}_{n,h_n}(\mathbf{x}) - \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x})|.$$

We can therefore write, for any  $n \geq 1$ , the inequalities

$$\begin{aligned}
|\Delta_{1,n,h_n,\beta}| &= \left| \int_{A_{n,\beta}} \left\{ \log \hat{f}_{n,h_n}(\mathbf{x}) - \log \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) \right\} \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) d\mathbf{x} \right| \\
&\leq \int_{A_{n,\beta}} \left| \log \hat{f}_{n,h_n}(\mathbf{x}) - \log \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) \right| \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) d\mathbf{x} \\
&\leq \frac{2}{(\log n)^{-\beta}} \int_{A_{n,\beta}} |\hat{f}_{n,h_n}(\mathbf{x}) - \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x})| \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) d\mathbf{x} \\
&\leq \frac{2}{(\log n)^{-\beta}} \sup_{\mathbf{x} \in A_{n,\beta}} |\mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) - \hat{f}_{n,h_n}(\mathbf{x})| \int_{A_{n,\beta}} \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) d\mathbf{x} \\
&\leq \frac{2}{(\log n)^{-\beta}} \sup_{\mathbf{x} \in \mathbb{R}^d} |\mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) - \hat{f}_{n,h_n}(\mathbf{x})| \int_{\mathbb{R}^d} \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) d\mathbf{x}.
\end{aligned}$$

We note that

$$\left| \int_{\mathbb{R}^d} \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) d\mathbf{x} - \int_{\mathbb{R}^d} f(\mathbf{x}) d\mathbf{x} \right| \leq \int_{\mathbb{R}^d} |\mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) - f(\mathbf{x})| d\mathbf{x}. \quad (8)$$

By combining (8) with Theorem 9.1, page 79, in Devroye and Lugosi (2001), we obtain that

$$\lim_{h \rightarrow 0} \int_{\mathbb{R}^d} \mathbb{E} \hat{f}_{n,h}(\mathbf{x}) d\mathbf{x} = \int_{\mathbb{R}^d} f(\mathbf{x}) d\mathbf{x} = 1.$$

Since  $h_n \leq b_n$  and  $b_n \downarrow 0$  as  $n \rightarrow \infty$ , there exists a positive constant  $C_1$  such that for all  $n$  sufficiently large

$$|\Delta_{1,n,h_n,\beta}| \leq \frac{C_1}{(\log n)^{-\beta}} \sup_{\mathbf{x} \in \mathbb{R}^d} |\mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) - \hat{f}_{n,h_n}(\mathbf{x})|. \quad (9)$$

We next evaluate the second term  $\Delta_{2,n,h_n,\beta}$  in the right side of (7). Since  $|\log z| \leq \frac{1}{z} + z$ , for all  $z > 0$ , we see that

$$\begin{aligned}
|\Delta_{2,n,h_n,\beta}| &= \left| \int_{A_{n,\beta}} \left\{ \hat{f}_{n,h_n}(\mathbf{x}) - \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x}) \right\} \log \hat{f}_{n,h_n}(\mathbf{x}) d\mathbf{x} \right| \\
&\leq \int_{A_{n,\beta}} |\hat{f}_{n,h_n}(\mathbf{x}) - \mathbb{E} \hat{f}_{n,h_n}(\mathbf{x})| \left[ \frac{1}{\hat{f}_{n,h_n}(\mathbf{x})} + \hat{f}_{n,h_n}(\mathbf{x}) \right] d\mathbf{x}.
\end{aligned}$$

Similarly as above, we get, for any  $\mathbf{x} \in A_{n,\beta}$ ,

$$\begin{aligned} \frac{1}{\hat{f}_{n,h_n}(\mathbf{x})} + \hat{f}_{n,h_n}(\mathbf{x}) &= \left( \frac{1}{\hat{f}_{n,h_n}(\mathbf{x})\hat{f}_{n,h_n}(\mathbf{x})} + 1 \right) \hat{f}_{n,h_n}(\mathbf{x}) \\ &\leq \left( \frac{1}{(\log n)^{-2\beta}} + 1 \right) \hat{f}_{n,h_n}(\mathbf{x}). \end{aligned}$$

We can therefore write, for any  $n \geq 1$ ,

$$\begin{aligned} |\Delta_{2,n,h_n,\beta}| &\leq \left( \frac{1}{(\log n)^{-2\beta}} + 1 \right) \int_{A_{n,\beta}} |\mathbb{E}\hat{f}_{n,h_n}(\mathbf{x}) - \hat{f}_{n,h_n}(\mathbf{x})| \hat{f}_{n,h_n}(\mathbf{x}) d\mathbf{x} \\ &\leq \left( \frac{1}{(\log n)^{-2\beta}} + 1 \right) \sup_{\mathbf{x} \in A_{n,\beta}} \left| \mathbb{E}\hat{f}_{n,h_n}(\mathbf{x}) - \hat{f}_{n,h_n}(\mathbf{x}) \right| \int_{A_{n,\beta}} \hat{f}_{n,h_n}(\mathbf{x}) d\mathbf{x} \\ &\leq \left( \frac{1}{(\log n)^{-2\beta}} + 1 \right) \sup_{\mathbf{x} \in A_{n,\beta}} \left| \mathbb{E}\hat{f}_{n,h_n}(\mathbf{x}) - \hat{f}_{n,h_n}(\mathbf{x}) \right| \int_{\mathbb{R}^d} \hat{f}_{n,h_n}(\mathbf{x}) d\mathbf{x}. \end{aligned}$$

Observe that

$$\left| \int_{\mathbb{R}^d} \hat{f}_{n,h_n}(\mathbf{x}) d\mathbf{x} - \int_{\mathbb{R}^d} f(\mathbf{x}) d\mathbf{x} \right| \leq \int_{\mathbb{R}^d} |\hat{f}_{n,h_n}(\mathbf{x}) - f(\mathbf{x})| d\mathbf{x}. \quad (10)$$

By combining (10) with Theorem 3.1, page 30, in Devroye (1987), we conclude that, almost surely

$$\int_{\mathbb{R}^d} \hat{f}_{n,h_n}(\mathbf{x}) d\mathbf{x} \rightarrow \int_{\mathbb{R}^d} f(\mathbf{x}) d\mathbf{x} = 1 \quad \text{as } h_n \rightarrow 0 \text{ and } nh_n \rightarrow \infty.$$

Since  $cn^{-1} \log n \leq h_n \leq b_n$  and  $b_n \downarrow 0$  as  $n \rightarrow \infty$ , there exists a positive constant  $C_2$  such that, almost surely for  $n$  sufficiently large,

$$|\Delta_{2,n,h_n,\beta}| \leq \left( \frac{1}{(\log n)^{-2\beta}} + 1 \right) C_2 \sup_{\mathbf{x} \in \mathbb{R}^d} \left| \mathbb{E}\hat{f}_{n,h_n}(\mathbf{x}) - \hat{f}_{n,h_n}(\mathbf{x}) \right|. \quad (11)$$

We now impose some slightly more general assumptions on the kernel  $K(\cdot)$  than that of Theorem

1.1. Consider the class of functions

$$\mathcal{K} = \left\{ K((\mathbf{x} - \cdot)/h^{1/d}) : h > 0, \mathbf{x} \in \mathbb{R}^d \right\}.$$

For  $\varepsilon > 0$ , set  $N(\varepsilon, \mathcal{K}) = \sup_Q N(\kappa\varepsilon, \mathcal{K}, d_Q)$ , where the supremum is taken over all probability measures  $Q$  on  $(\mathbb{R}^d, \mathcal{B})$ . Here,  $d_Q$  denotes the  $L_2(Q)$ -metric and  $N(\kappa\varepsilon, \mathcal{K}, d_Q)$  is the minimal number of balls  $\{g : d_Q(g, g') < \varepsilon\}$  of  $d_Q$ -radius  $\varepsilon$  needed to cover  $\mathcal{K}$ . We assume that  $\mathcal{K}$  satisfies the following uniform entropy condition.

**(K.5)** for some  $C > 0$  and  $\nu > 0$ ,

$$N(\varepsilon, \mathcal{K}) \leq C\varepsilon^{-\nu}, 0 < \varepsilon < 1. \quad (12)$$

Finally, to avoid using outer probability measures in all of statements, we impose the following measurability assumption.

**(K6)**  $\mathcal{K}$  is a pointwise measurable class, that is, there exists a countable subclass  $\mathcal{K}_0$  of  $\mathcal{K}$  such that we can find for any function  $g(\cdot) \in \mathcal{K}$  a sequence of functions  $\{g_m(\cdot) : m \geq 1\}$  in  $\mathcal{K}_0$  for which

$$g_m(\mathbf{z}) \longrightarrow g(\mathbf{z}), \quad \mathbf{z} \in \mathbb{R}^d.$$

Remark that Condition (K.5) is satisfied whenever  $K(\cdot)$  is of bounded variation, and Condition (K.6) is satisfied whenever  $K(\cdot)$  is right continuous (refer to Deheuvels and Mason (2004) and Einmahl and Mason (2005)).

By Theorem 1 of Einmahl and Mason (2005), whenever  $K(\cdot)$  is measurable and satisfies (K3)-(K6), and when  $f(\cdot)$  is bounded, we have for each  $c > 0$ , and for a suitable function  $\Sigma(c)$ , with probability 1,

$$\limsup_{n \rightarrow \infty} \sup_{cn^{-1} \log n \leq h \leq 1} \frac{\sqrt{nh} \|\hat{f}_{n,h} - \mathbb{E}\hat{f}_{n,h}\|_\infty}{\sqrt{\log(1/h) \vee \log \log n}} = \Sigma(c), \quad (13)$$

which implies, in view of (9) and (11), that, with probability 1,

$$\limsup_{n \rightarrow \infty} \sup_{cn^{-1}(\log n)^{1+4\beta} \leq h \leq b_n} \frac{\sqrt{nh} |\Delta_{1,n,h,\beta}|}{\sqrt{\{\log n\}^{4\beta} (\log(1/h) \vee \log \log n)}} = 0, \quad (14)$$

and

$$\limsup_{n \rightarrow \infty} \sup_{cn^{-1}(\log n)^{1+4\beta} \leq h \leq b_n} \frac{\sqrt{nh} |\Delta_{2,n,h,\beta}|}{\sqrt{\{\log n\}^{4\beta} (\log(1/h) \vee \log \log n)}} \leq \Upsilon(c). \quad (15)$$

Recalling (7), the proof of Theorem (1.1) is completed by combining (14) with (15).  $\square$

**Proof of Corollary 1.2.** Recall  $A_{n,\beta} = \{\mathbf{x} : \hat{f}_{n,h_n}(\mathbf{x}) \geq (\log_+ n)^{-\beta}\}$  and let  $A_{n,\beta}^c$  the complement of  $A_{n,\beta}$  in  $\mathbb{R}^d$  (i.e.  $A_{n,\beta}^c = \{\mathbf{x} : \hat{f}_{n,h_n}(\mathbf{x}) < (\log_+ n)^{-\beta}\}$ ). We repeat the arguments above with the formal change of  $H_{n,h_n,\beta}(f)$  by  $H(f)$ . We show that there exists positive constants  $D_1$  and  $D_2$  such that, for all  $n$  sufficiently large,

$$\begin{aligned} |\hat{\mathbb{E}}H_{n,h_n,\beta}(f) - H(f)| &\leq \left| \int_{A_{n,\beta}^c} f(\mathbf{x}) \log(f(\mathbf{x})) d\mathbf{x} \right| + \frac{D_1}{(\log n)^{-\beta}} \sup_{\mathbf{x} \in \mathbb{R}^d} |\mathbb{E}\hat{f}_{n,h_n}(\mathbf{x}) - f(\mathbf{x})| \\ &\quad + \left( \frac{1}{(\log n)^{-2\beta}} + 1 \right) D_2 \sup_{\mathbf{x} \in \mathbb{R}^d} |\mathbb{E}\hat{f}_{n,h_n}(\mathbf{x}) - f(\mathbf{x})|. \end{aligned} \quad (16)$$

We know (see, e.g., Einmahl and Mason (2005)), that when the density  $f(\cdot)$  is uniformly Lipschitz and continuous, we have for each  $a_n < h < b_n$ , as  $n \rightarrow \infty$ ,

$$\|\mathbb{E}\hat{f}_{n,h}(\mathbf{x}) - f(\mathbf{x})\|_\infty = O(b_n^{1/d}).$$

Thus, we have

$$\lim_{n \rightarrow \infty} \sup_{a_n \leq h \leq b_n} (\log n)^{2\beta} \|\mathbb{E}\hat{f}_{n,h}(\mathbf{x}) - f(\mathbf{x})\|_\infty = 0.$$

This when combined with (16), entails that, as  $n \rightarrow \infty$ ,

$$\sup_{a_n \leq h \leq b_n} \|\hat{\mathbb{E}}H_{n,h,\beta}(f) - H(f)\| \rightarrow 0. \quad (17)$$

Using (17) in connection with (5) imply (6).  $\square$

## References

- Akaike, H. (1954). An approximation to the density function. *Ann. Inst. Statist. Math., Tokyo*, **6**, 127–132.
- Beirlant, J., Dudewicz, E. J., Györfi, L., and van der Meulen, E. C. (1997). Nonparametric entropy estimation: an overview. *Int. J. Math. Stat. Sci.*, **6**(1), 17–39.
- Bosq, D. and Lecoutre, J.-P. (1987). *Théorie de l'estimation fonctionnelle*. Économie et Statistiques Avancées. Economica, Paris.
- Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition.
- Deheuvels, P. (2000). Uniform limit laws for kernel density estimators on possibly unbounded intervals. In *Recent advances in reliability theory (Bordeaux, 2000)*, Stat. Ind. Technol., pages 477–492. Birkhäuser Boston, Boston, MA.
- Deheuvels, P. and Mason, D. M. (2004). General asymptotic confidence bands based on kernel-type function estimators. *Stat. Inference Stoch. Process.*, **7**(3), 225–277.
- Devroye, L. (1987). *A course in density estimation*, volume 14 of *Progress in Probability and Statistics*. Birkhäuser Boston Inc., Boston, MA.
- Devroye, L. and Györfi, L. (1985). *Nonparametric density estimation*. Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics. John Wiley & Sons Inc., New York. The  $L_1$  view.
- Devroye, L. and Lugosi, G. (2001). *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York.
- Dony, J. and Einmahl, U. (2006). Weighted uniform consistency of kernel density estimators with general bandwidth sequences. *Electron. J. Probab.*, **11**, no. 33, 844–859 (electronic).

- Einmahl, U. and Mason, D. M. (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. *J. Theoret. Probab.*, **13**(1), 1–37.
- Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.*, **33**(3), 1380–1403.
- Györfi, L. and van der Meulen, E. C. (1991). On the nonparametric estimation of the entropy functional. In *Nonparametric functional estimation and related topics (Spetses, 1990)*, volume 335 of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.*, pages 81–95. Kluwer Acad. Publ., Dordrecht.
- Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, **33**, 1065–1076.
- Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, **27**, 832–837.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.*, **27**, 379–423, 623–656.